UNITED NATIONS

E

Economic and Social Council

Distr.
GENERAL

ECE/CES/AC.6/2008/6
3 March 2008

Original: ENGLISH

ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/Eurostat Meeting on Population and Housing Censuses

Eleventh Meeting
Geneva, 13-15 May 2008
Item 2 of the provisional agenda

CENSUS QUALITY ASSURANCE AND EVALUATION

Quality assessment of the register-based Slovenian census 2011

Note by the Statistical Office of the Republic of Slovenia

Summary

The use of the administrative data as a means for the costs and burden reduction is a very popular practice in the modern official statistics. In Slovenia already the 2002 Census was partly based on administrative data sources, while the next, 2011 Census is planned to be fully register-based. The quality assessment of the census processes and results is an important obligation of statistical offices. In the document some considerations are presented about the modifications of the quality assessment concepts when moving from a conventional to a register-based census. The first part of the document is devoted to some general reflections, while in the second part the focus is on the plans for the quality assessment of the Slovenian 2011 register-based census.

I. INTRODUCTION

1. The use of administrative data, especially data kept in the registers of different types, is becoming more and more popular in different fields of the official statistics. Since administrative data have already been used for a long time in some parts of the statistical processes such as sampling frame construction process or calibration procedures, in recent years this use has rapidly been widened to the use of administrative data as a direct data source. The housing and population census as probably the most costly and the most burdensome statistical survey is of course no exception in these processes. In the last decade a vast number of activities were conducted in different statistical offices aiming at developing the adequate environment and efficient processes for compiling as much as possible census data from the already existing administrative sources. The leading role in these processes certainly belongs to the Nordic countries, where the process of moving from a conventional to a completely register-based census has almost finished.

2. The Statistical Office of the Republic of Slovenia (hereinafter SORS) has been following the Nordic model of wide and comprehensive use of administrative data for a long time. Many administrative registers that are now under maintenance of other authorities have been set-up and for some time also maintained by SORS. This fact presents a significant advantage for SORS, since there is still a deep knowledge of the conceptual syntax of these registers at the office and it is natural that these registers are still widely used for different statistical purposes. While the last Slovenian Census of Population and Housing was still a combined one, based on combining data from registers and from conventional field surveys, the next one in 2011 is already planned to be fully register-based. At this stage some quite demanding challenges are still being faced, which have to be overcome as a precondition for the implementation of such a census, but it is hoped that these problems will be successfully solved and that the ambitious plan will be fulfilled.

3. One of the important obligations and responsibilities of data producers is that they regularly and systematically monitor and report the quality of their products. Eurostat, together with the Member States, has built a widely accepted conceptual framework for the quality assessment and reporting which is based on six quality components (relevance, accuracy, timeliness and punctuality, accessibility and clarity, comparability, and coherence) complemented with the costs and burden as an additional component. The fact is that this framework with the underlying concepts has been mostly defined for the purposes of the classical statistical surveys. Many authors have recently pointed out that there is still a lack of theoretical framework for the systematic quality assessment of the register-based statistics and this is surely a challenge for the near future.

4. In the document some considerations are presented about the quality assessment and reporting for the next (hopefully completely) register-based census. While the quality report for the 2002 Slovenian Census of Population and Housing (which was publicly disseminated on SORS's website) still more or less followed the classical quality report syntax, the main focus of the quality assessment and reporting for the 2011 Census will certainly be moved to some other quality issues.

5. In the first part of the document some general considerations will be presented about the quality assessment of the register-based statistics, while in the second part the focus will be on the Slovenian reality. Since SORS is still at the early preparation phase for the next census, also the considerations about the quality assessment and reporting concepts are in the formation phase and they will certainly be a subject of further examinations.

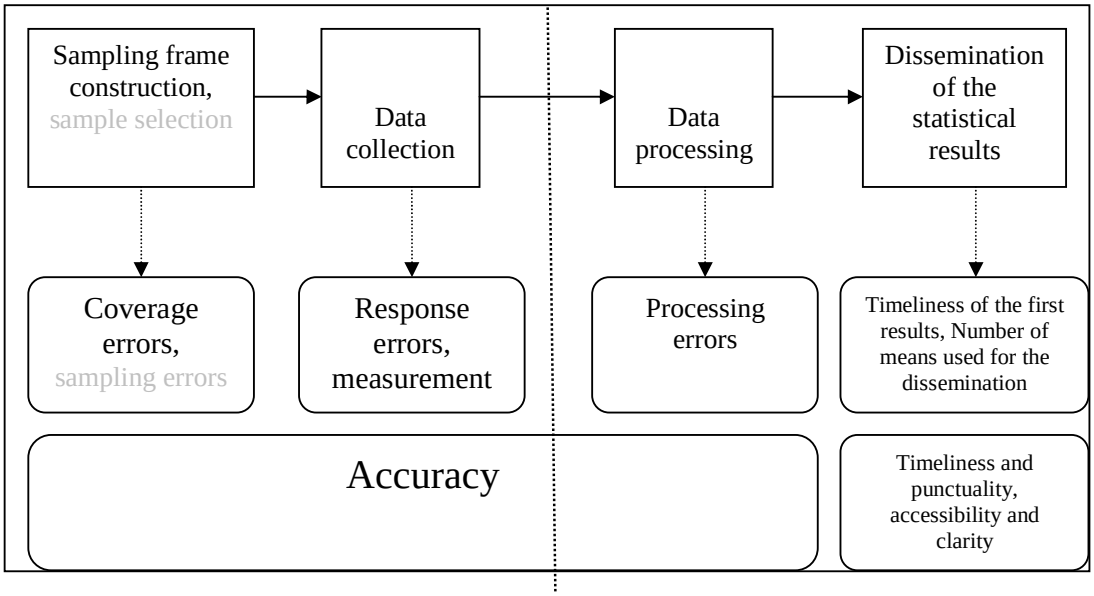
II. QUALITY ASSESSMENT: FROM CLASSICAL TO REGISTER-BASED STATISTICS

6. The modern quality assessment model considers the statistical product as a result of a complex statistical process and when assessing the quality of the statistical product it is of great importance to monitor and assess different parts of the process as well as the final statistical product. The general quality assessment framework, which has been in the last decade widely accepted inside the European statistical system, provides a strong tool for standard and consistent activities in this field.

7. There are two main concepts that are provided in this framework: standard quality components and for each component a set of standard quality indicators. There is a variety of different models by which these concepts could be presented, each of them using certain degree of simplification. For the purpose of this document six quality components will firstly be divided into two parts. The first part will consist of components which consider the quality of the product itself and the second part will consist of the components more focused on the quality of the statistical process.

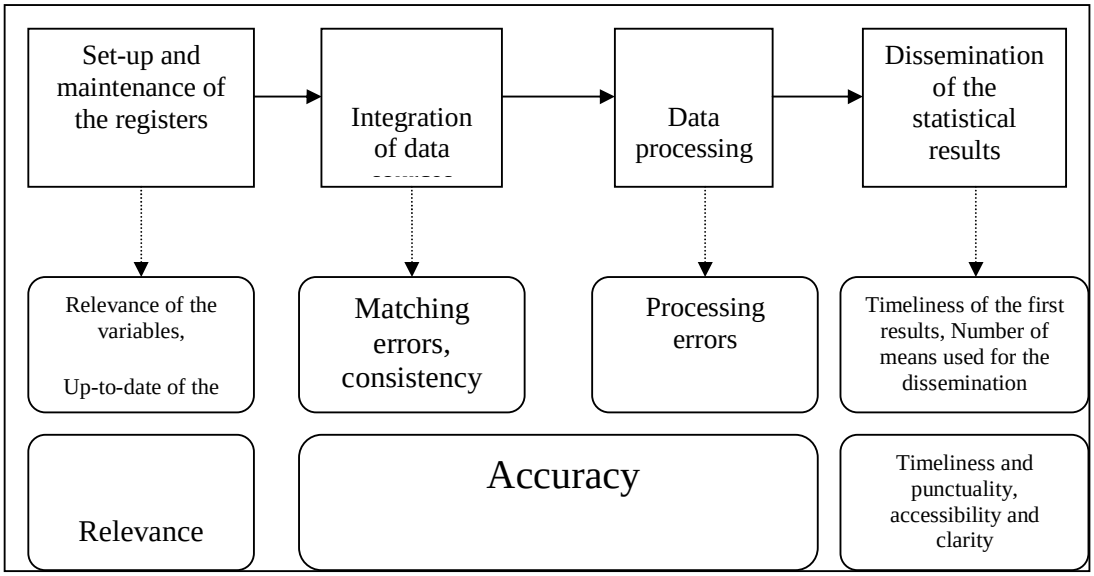
8. The “product oriented” components are: relevance; coherence; comparability. The main characteristic of these components is the fact that for their assessment it is not sufficient to monitor only the statistical process but there is also need for some information from the wider environment, such as user needs, user satisfaction, coherence with some other statistics, etc. One further characteristic of these components is that there is sometimes difficult to numerically assess the degree of quality and that there are not many really widely used quality indicators derived from these components.

9. The “process oriented” components are: accuracy; timeliness and punctuality; accessibility and clarity. These components could more or less be assessed by monitoring different parts of the statistical process. The process of the “classical” survey could roughly be divided into four consecutive steps. In the picture the (simplified) flow of the process is presented, where in each step some of the most typical quality indicators, used for the quality assessment, are given.



10. There is not much difference when applying such a model for a sample survey or for a census. In fact the only general difference is that the "shadowed" parts in the picture don't apply in the case of the census, while the other components could be equally considered in both types of surveys.

11. Now, how this model could be adopted when moving from a conventional to a register-based census? Since such a change of the approach at least partly demands different view to the whole process, it most of all influences those parts of the process which are in the picture positioned left of the dashed horizontal line. Since classical data collection, designed especially for the statistical purposes, is replaced with the integration of the data from different registers, the quality assessment model could now be presented as follows:



12. The one important difference between the two approaches is that the relevance becomes one of the "process oriented" components in the second case. In fact this component is in the case of a register-based census the basic source which could influence most of the other quality components.

13. Now, what are the most important issues that someone should take into account when the classical survey quality model is adopted for the purposes of a register-based census?

(a) The process of the classical survey data collection is designed especially for the statistical purposes while on the other hand the data from registers are mostly not primarily intended for statistical use. Therefore, the detailed investigation about the relevance of each of the variables is of great importance. The special attention should be given to those variables whose values are derived from different sources.
(b) Each of the registers used in the register-based census has its own history, its own administrative environment, its own way of data collection and processing, etc. So, it is important that in the preparation phase these characteristics of each register are carefully studied and documented. At least the overview of these findings should be put in the quality report.
(c) The legal acts and administrative regulations can have an important influence on the quality of the incoming data. Therefore, a part of the quality report should be devoted to these issues.
(d) In the case of a conventional census, the collected data usually refer to the same critical time point (reference date and hour). In a register-based census it is difficult to reach such coherence in the reference time periods of the data from the different sources. Hence, the cases of the departures from the common reference time-point should be documented and reported in the quality report.

14. So far not much has been said about the additional quality component *costs and burdens*. Although the component itself shouldn't change much when moving from a conventional to a register-based census, the real difference should come with the numerical assessments of this component. By default the costs and especially the burdens should lower considerably in the case of a register-based census. Therefore, the costs and burdens assessment together with its comparison to the classical survey's costs and burdens should also be (an important part) of the quality report.

III. QUALITY PLANS FOR THE SLOVENIAN REGISTER-BASED CENSUS

A. Short background and available administrative and statistical sources for Slovenia's first register-based census in 2011

15. With a register-based census the 'census data are produced using the method of register estimation, in which several register sources are simultaneously used to define for each statistical unit the value of the relevant variable'[1]. At Slovenian 2011 register-based census data from three

[1] *CES Recommendations for the 2010 Censuses of Population and Housing*. United Nations, New York and Geneva, 2006, p. 158.

main administrative sources will be used; from the Central Population Register (hereinafter CPR), from the e-Database of Households and from the Register of Dwellings. These sources will be used in order to define the two target populations; the population of persons and the population of dwellings. In addition, other administrative and statistical sources will be used with the aim to cover obligatory and nationally interesting census topics. Other available sources are: the Register of Employment, the Register of Unemployment, the Register of Foreigners, the Register of Territorial Units, the Tax Database, social statistics surveys, etc. In some cases (mostly for the household and family structure) also data from the 2002 Census will be used. Data from other data sources will be linked to the target populations in the data integration phase.

16. The main obstacles which the Slovenian NSI is facing at the moment concerning the data sources for the register-based census are:

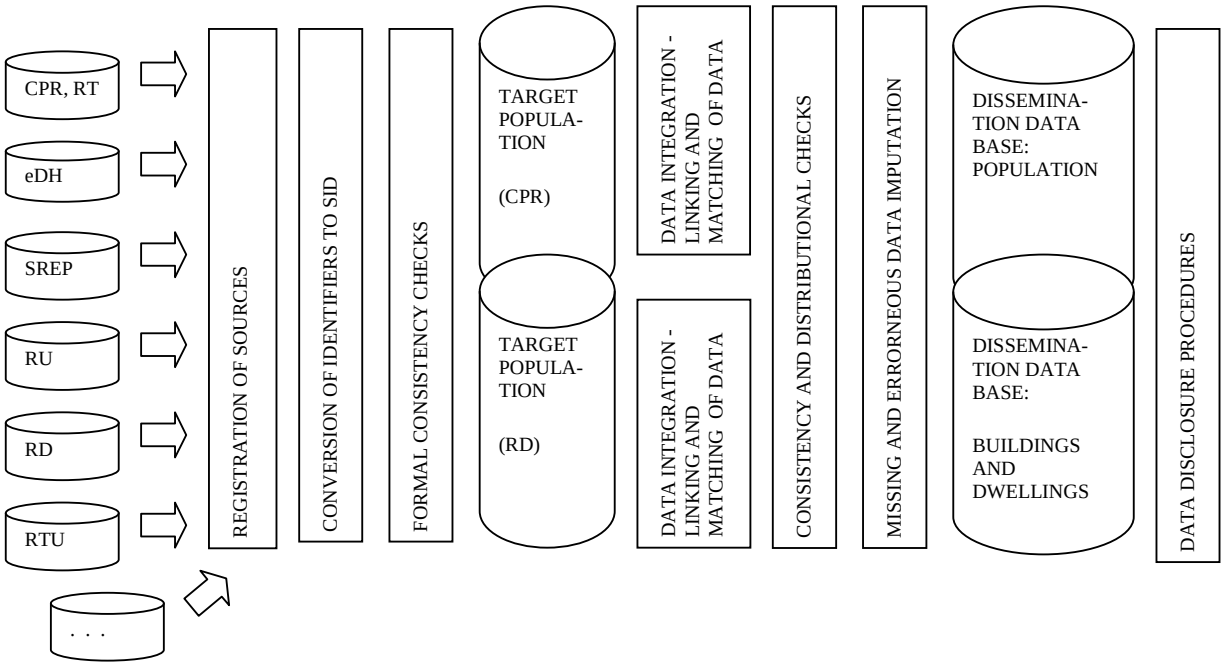
(a) The Register of Dwellings is still in the establishment phase (managed by the Surveying and Mapping Authority of the Republic of Slovenia);
(b) The quality of the data in the eDatabase of Households is not satisfactory (maintained by the Ministry of the Interior) since it has been just recently established (computerized);
(c) Data on relationships among persons in the CPR are not as complete as they should be so that they would enable SORS to easily identify all relationships among persons in the families.

17. If the Slovenian situation is compared to the situation in the Nordic countries, the biggest difference is that in the CPR some relationships between the family members exist, but only for the vital events such as births, marriages, etc., that took place after April 1988. The information on relationships in the family is not so complete as to enable simple forming of families. Families could be formed on the basis of data on age, marital status and gender of persons residing together in a dwelling unit, with the precondition that a dwelling number is available in the CPR. In addition, for some population groups also data on the relationships in the family from the 2002 Census could be taken (for example, for people who have not moved since the last Census).

18. Of course, there is still an option to introduce an ‘ad hoc’ module in one of the largest household surveys, which would enable to collect information on the families. However, these data could be used only to asses the completeness of the coverage and quality of the data obtained from the administrative and statistical sources.

B. Process and quality assessment foreseen for the 2011 register-based census

19. At SORS the following process of a register-based census is foreseen:



CPR - the Central Population Register

RT – the Register of Foreigners

eDH – the eDatabase of Households

SREP – the Statistical Register of Employed Persons

RU – the Register of Unemployed

RD – the Register of Dwellings

RTU – the Register of Territorial Units

... – other sources

20. The first step in the register-based system will be an identification of the relevant administrative and statistical sources. Partly this task has been already covered at the 2002 Census of the Population and Housing, but since then many data were computerized and many new administrative sources were established - mostly within the Ministry of the Interior and the Ministry of Public Administration. Concerning the quality assessment, the component of relevance will be covered already in this stage.

21. For each source entering the register-based census system the following information will be documented (on the basis of a template):

(a) name of the source;

(b) description of the source:

(i) date of the establishment,
(ii) legal basis,
(iii) identifiers used,
(iv) system of updating,
(v) structure,

(c) method and frequency of acceptance of the source at SORS;
(d) data editing at SORS.

22. In each source relevant variables for the 2011 Census will be identified. For the selected relevant variables the time frame and deviations from the reference date (and hour) will be studied and documented. If there are several sources for the same relevant variable, the priority list of sources will be set.

23. On the basis of documentation, each source will go through the registration, conversion of the identifiers to statistical identifier (to a so-called SID) and through the validity checks process.

24. It is necessary that each source is registered, but at that stage the only quality measure which could be implemented is whether it was possible to register the source or not. At conversion of the identifiers to statistical identifier the number of inadequate records for the conversion could be measured. On that basis the number of records (within the source and in all sources together entering in the register-based census system) which passed through to the following stages of the process could be assessed. With the validity checks the consistency with the predefined range of variables in the source will be checked. Some obvious outliers will be also identified at this stage of the process.

25. After validity controls and on the basis of rules set (according to the definitions agreed at the international level) the two basic target populations will be set – the population of persons and the population of dwellings. Since the CPR data are, at least if the coverage of persons present on the territory of Slovenia is considered, of a high quality, the definition of usual residents should not be problematic. The only inconsistencies that could be expected are when the CPR units will be linked to the units from the Register of Foreigners. Some duplicates can occur due to the different purpose of the maintenance of those two registers. At present not much is known on the coverage of the target population in the Register of Dwellings since the mentioned register is in the establishment phase. But the register will be based on the Census of Real Estate which took place in 2006 and the under coverage could be foreseen only for the newly built buildings and dwellings.

26. The phase of the data integration could undoubtedly be considered as the crucial stage of the statistical process when quality issues are concerned. While the first part of the quality report, devoted to the relevance of the registers and variables, will more or less be descriptive and textual, the “data integration part” should much more be based on the numerical information. Since this part of the process hasn’t yet been covered in the “classical” quality components, there is a clear lack of the standard quality indicators for this phase. It is believed that these issues will be broadly discussed in the near future; here two suggestions for the indicators that could be used are presented.

27. In the integration process several different sources will be merged to a target population list. This merging is planned to be accomplished through the two consecutive steps. In the first step the part of the data source will be merged by using the direct matching by the unique identifier. For the remaining part, which will not be successfully merged through the direct matching process, the statistical matching will be used. Monitoring this process, three matching rates can be defined. The overall matching rate provides the rate of all the units (regarding the target population) for which the data from the administrative source have successfully been merged (directly or statistically). The direct matching rate provides the rate of the units merged by direct matching and the statistical matching rate the rate of the units merged by statistical matching.

28. For the variables whose values will be derived from different sources, the consistency of the values from different sources can be measured. If the first priority variable source is labeled Y_R and the alternative variable source is labeled Y_A, the consistency rate can be defined as the ratio between the number of units for which the condition $|Y_R - Y_A|/Y_R < p$ (e.g. $p = 0.01$) holds and the number of all the units for which the data are available from both sources.

29. Besides the part of the QR which will be devoted to the direct monitoring of the statistical process, attention will naturally be paid also to the remaining quality components: comparability and coherence. According to the present plans, the comparability component will mostly be devoted to the comparison of the results of the 2011 register-based census with the results of the 2002 (mostly) conventional census. Special attention will be put on the investigation of the population structure by the enumeration areas. Since it is known that the administrative residence of the persons many times differs from the actual (usual) residence, some significant differences can eventually be found in these data.

30. The main question which should come with the coherence component should be: “Are the results of the register-based survey coherent with the results of some other, “classical” statistical surveys such as EU-SILC, LFS and HBS?” Hence, the main part of the quality report coherence component should be devoted to the presentation and description of the largest differences between these results. The goal is to give the user a clear picture about the consequences that the difference in the two approaches could cause.

IV. CONCLUSIONS

31. Most of the National Statistical Instîtutes within the European statistical system have already started intensive work on the preparations for the next 2010 population and housing census round. Following the “Nordic example”, most of the countries are planning to use as much as possible of already existing administrative data in the collection phase of the census. Slovenia is one of the countries that are planning to carry out a fully register-based census.

32. Among important issues which should certainly be taken into account in the preparation phase of a census are the plans related to the foreseen quality assessment of the census process and results. Since the existing quality concepts, which have been developed in the European statistical system in the last decade, were mostly tailored for the classical statistical surveys, considerations about some new “quality approaches” are needed. In the paper some considerations were presented about this topic. In the first part the focus was on the general

theoretical framework, while in the second part it was on the specific situation in Slovenia. Here some of the most important issues are emphasized:

(a) The movement from a conventional to a register-based statistical survey demands an adjusted approach in the data quality assessment. Most significant changes come out of the assessment of the survey preparation and data collection processes.
(b) The relevance of the administrative and statistical sources and the relevance of the statistical variables, derived from these sources, is the crucial component of the new approach.
(c) The exhaustive documentation which should describe the different aspects of the incoming sources and their administrative environment should become an important part of the quality report.
(d) Also the set of standard quality indicators should be adopted for the case of register-based survey. Most of the new indicators should describe the quality of the data integration phase.
(e) The main goal of the comparability and coherence component should be that they would enable the user an insight into the differences which are caused by the register-based approach.

REFERENCES

CES Recommendations for the 2010 Censuses of Population and Housing. United Nations, New York and Geneva, 2006.

Biemer Paul B., Lyberg Lars E.: Introduction to Survey Quality: John Wiley & sons, 2003.

Methodological documents, Handbook “How to make quality report: Working Group “Assessment of quality in statistics”, Sixth meeting, Luxembourg 2-3 October 2003.

Methodological documents, Standard Report: Working Group “Assessment of quality in statistics”, Sixth meeting, Luxembourg 2-3 October 2003.

Oblak Flander, A., 2007. Opportunities and Challenges of a Register-Based Census of Population and Housing – the Case of Slovenia. (Unpublished. Document prepared for the Seminar on Registers in Statistics - methodology and quality, 21–23 May, 2007, Helsinki)

Standard Quality Indicators - Producer oriented: Working Group “Assessment of quality in statistics”, Sixth meeting, Luxembourg 2-3 October 2003.

Wallgren A., Wallgren B.: “Register-based Statistics; Administrative Data for Statistical Purposes”: John Wiley & sons, 2007.
